NASA/TM—2000-210381

Sub-Nyquist Distortions in Sampled Data, Waveform Recording, and Video Imaging

Glenn L. Williams
Glenn Research Center, Cleveland, Ohio

National Aeronautics and
Space Administration

Glenn Research Center

October 2000

This report contains preliminary findings, subject to revision as analysis proceeds.

Available from

NASA Center for Aerospace Information
7121 Standard Drive
Hanover, MD 21076
Price Code: A03

National Technical Information Service
5285 Port Royal Road
Springfield, VA 22100
Price Code: A03

Available electronically at http://gltrs.grc.nasa.gov/GLTRS

SUB-NYQUIST DISTORTIONS IN SAMPLED DATA, WAVEFORM RECORDING, AND VIDEO IMAGING

Glenn L. Williams
National Aeronautics and Space Administration
Glenn Research Center
Cleveland, Ohio 44135

SUMMARY

Investigations of aliasing effects in digital waveform sampling have revealed the existence of a mathematical field and a pseudo-alias domain lying to the left of a "Nyquist line" in a plane defining the boundary between two domains of sampling. To the right of the line lies the classic alias domain. For signals band-limited below the Nyquist limit, displayed output may show a false modulation envelope whenever inadequate signal reconstruction is used before display. The effect occurs whenever the sample rate and the signal frequency are related by ratios of mutually prime integers. For cost and technical reasons, the waveform display devices omit the required reconstruction steps. Belying the principal of a 10:1 sampling ratio being "good enough," this distortion easily occurs in graphed one-dimensional waveforms and two-dimensional images and occurs daily on television.

1. INTRODUCTION

An instructive experiment begins by coupling the sine-wave output of an analog signal generator into an analog strip-chart recorder and plotting the trace of the waveform, as in figure 1(a). Then, without increasing the chart paper speed, the frequency of the sine-wave signal is slowly increased until the displayed signal trace just merges into a solid band of color, as in figure 1(b). At some point, as the signal frequency is further increased, the painted band of signal recording will start to decrease in amplitude as the signal frequency starts to approach the upper limit of the frequency response of the recorder. This effect is shown in the remaining portions of figure 1. The roll-off in frequency response is, of course, caused by the frequency bandwidth limitations of the electronics (and sometimes the electromechanics) of the strip-chart recorder. Recorder manufacturers commonly print roll-off specifications in their instrument manuals, including the 3-decibel (3-db) down frequency point, where the recorder's displayed signal amplitude has dropped by a factor of one-half (defined as the bandwidth of the recorder).

Modern digital waveform recorders and oscilloscopes use extremely fast state-of-the-art analog-to-digital converters in the signal-coupling front end amplifiers, so that no longer is the analog low-pass frequency roll-off of any concern. A well-known principle of digitally sampled waveform acquisition is that the user must avoid the possibility of signal aliasing by restricting the input frequencies (including harmonics) to less than one-half the sampling frequency. This obeys the Nyquist-limit rule. But with analog-to-digital converters which can sample up to several hundred million times per second, there would seem to be no concern about having distorted waveform records in most everyday applications. The standard rule-of-thumb is just to avoid measuring signals containing frequencies greater than approximately one-tenth of the Nyquist limit (see below).

Inside the modern recorder, waveform playback is almost universally done by graphing vector or raster line approximations of the recorded waveform on a digital display, or for analog playback, by sending the data through a digital-to-analog converter, followed by resistive-capacitive (RC) or sin(x)/x filtering to smooth the analog output.

A half century of experience has iconified the Shannon (WKS) Sampling Theorem which prescribes an upper bound to the signal frequency such that (ref. 1):

> "If a function of time is limited to the band from 0 to W cycles per second, it is completely determined by giving its ordinates at a series of discrete points spaced 1/2W seconds apart in the manner indicated by the following result: If $f(t)$ has no frequencies over W cycles per second, then

$$f(t) = \sum_{n=-\infty}^{\infty} f\left(\frac{n}{2W}\right)\left(\frac{\sin \pi(2Wt - n)}{\pi(2Wt - n)}\right) \qquad (1)$$"

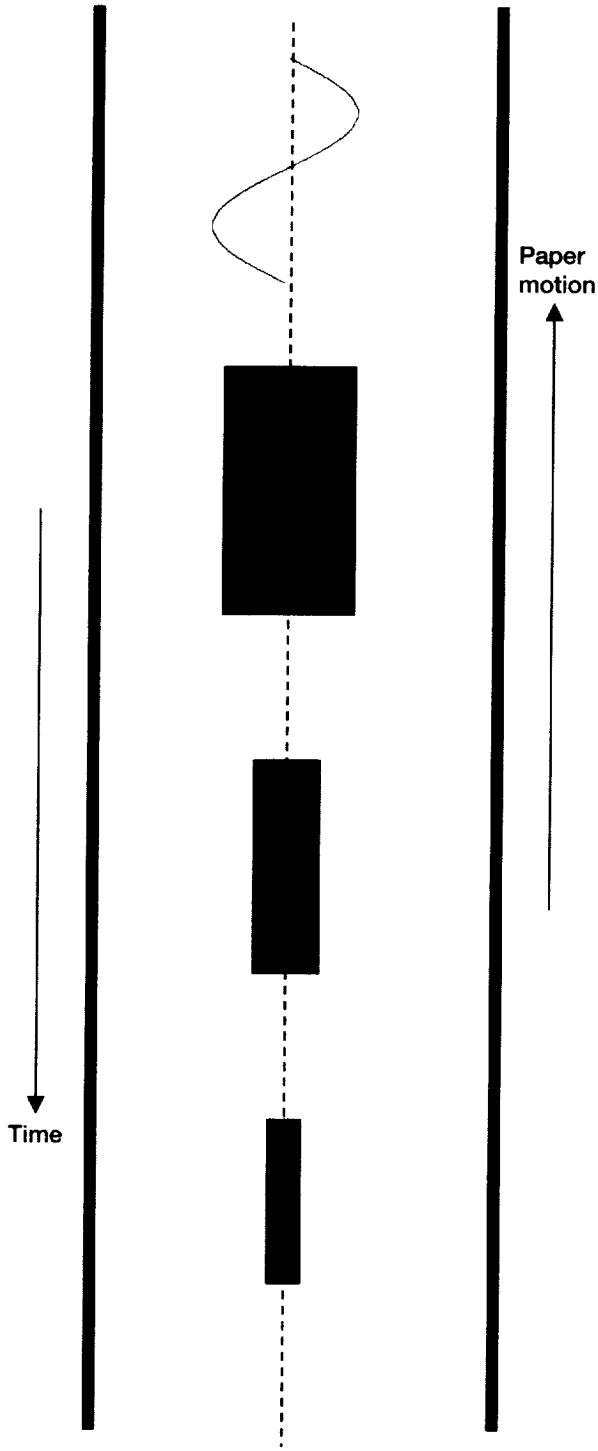

Figure 1.—(a) A slow sine wave signal is recorded on a strip chart recorder. (b) The signal frequency is increased until the trace turns into a solid band. (c) As the frequency continues to increase the amplitude of the solid band starts to roll off.

In countless articles, this theorem has been cited as justification for dropping concern about sampling distortion effects. All one needs to do is keep the signal frequency bandwidth well below one-half the sample frequency (which is also called the Nyquist frequency) and the "waveform can always be reconstructed completely because of Shannon's Theorem."

This report will show that neglecting the details of sampling theory can lead to serious misunderstandings about the appearance of displayed waveforms if certain special conditions exist. Serious distortions occur even when the input waveforms are correctly bandwidth limited or are even pure sine-waves. The distortions will be shown to occur at frequencies far below the sampling frequency and are not caused by frequency roll-off limitations in the analog-to-digital converter or elsewhere in the instrument. Rather, the distortions occur because most cost-effective signal waveform playback and display technologies leave out the complete waveform reconstruction required in accordance with the Sampling Integral.[1] As a result, when technical accuracy is of strategic importance, particular care must be applied to interpreting the display and reproduction of sampled-data waveforms on oscilloscopes, waveform recorders, spreadsheet charts and even television.

2. THEORY

In equation (1), with some manipulation of the mathematics, one also derives the "Nyquist rate" which is the minimum rate at which the signal needs to be sampled in order to reconstruct it completely. The Nyquist rate is one-half of the sample rate, or W, as expressed in the quotation above. Shannon's theorem can be paraphrased as "So long as the signal is band-limited to less than the Nyquist frequency, it can be completely reconstructed without any distortion."

Unfortunately this simple statement is often construed as implying that the waveform can be simply displayed without any other work. In actuality, the reconstruction requirements involve coping with mathematical and physical nuances, such as:

(a) the signal must be continuous forever in the past and in the future, in order to obtain all the sample points necessary to perform the reconstruction, and
(b) an infinitely fast digital processor is required to make all the computations in anything near real-time.

Given these ominous reconstruction requirements the user can hardly be blamed for ignoring reconstruction issues.

2.1 Definitions

Several definitions of terms are in order here. In this paper signal is usually the term applied to a continuous analog signal, as opposed to a discrete digital signal. Signals are processed by something of the nature of a digital sampling system in order to be translated into a table of numbers which can later be reproduced on some quasi-analog output means in order to portray the waveform as accurately as possible. A one-dimensional signal is a continuous analog signal as normally viewed on an oscilloscope or waveform recorder as a wavy line. A one-dimensional voltage signal as a function of time is represented by a continuous function $f(t)$ for the purposes of mathematical analysis. A two-dimensional signal is best thought of as a set of one-dimensional signals (raster lines with the amplitude shown in brightness level rather than y-axis displacement) representing thin slices of an image as viewed on a television or computer monitor, wherein the thin slices are ideally so thin that the eye sees only the composite image without seeing edges of the slices.

A distortionless representation of a signal on an oscilloscope screen or waveform recorder output means that except for a constant scaling factor, the amplitude along the displayed waveform always stands for and has the same "shape" as the original signal.

A sampled signal $f(t)$ is represented as a set of points in a table or a computer memory device which represent a very close approximation to the amplitude of the signal at the instants of time when the sampling device measured the original signal.

---

[1] See reference 3, p. 75. The discrete form of the integral is found in many texts, and appears as

$$f(t) = \sum_{n=-\infty}^{\infty} f(nT)\frac{\sin(\omega(t-nT))}{\omega(t-nT)}$$

Calculation of new data points (reconstruction) with this integral is so computationally intensive that systems designers are forced to leave out the process. Unfortunately, doing so immediately leads to the effects which are the topic of this report.

A band-limited signal contains only frequency components which lie between two end points of frequency, a high end point and a low end point. In almost all waveform recorders and oscillographs, the frequency band extends from zero frequency on the low end (a "dc" signal) to some upper ac signal frequency. Thus the instrument displaying the signal waveform could be thought of as a low-pass filter device. The concept of signal display applies equally well to other uses of the output signal information. One such other use is the audio output of a telephone circuit or the audio from a sound (music) system. Again, the output device seems to behave as a kind of low-pass filter.

2.2 Common Causes of Playback Distortion

Analog strip-chart recorders and oscilloscopes long ago became antiquated due to the improved technology and lower cost of digital waveform sampling, flat-panel display technology, and computers. Analog recorders and oscilloscopes have always suffered from high frequency roll-off of the electronic or electromechanical devices used to reproduce the waveform on the display media. High-speed signal digitization with an analog-to-digital converter eliminates the roll-off problem and the sampled data does not have to be reproduced in a manner which depends on the frequency limitations of analog output technology. The waveform samples can be stored in memory, transferred to disk or tape for long-term storage, or sent over computer networks. In those cases, display of the waveform data occurs in near-real time or well after the fact. As long as the visual appearance of the displayed waveform is pleasing and steady to the eye, the brain does not register the fact that the displayed or graphed waveform is made of tiny little segments.

A common practice used by manufacturers in their sales literature for digital oscilloscopes and waveform recorders is to list specifications claiming a "flat" frequency response. This claims is made due to the implicitly fast digital logic and extremely high sample rate used in the recorders. The claim of "flat response" is, again, a result of the common misinterpretation of the Shannon Sampling Theorem cited above. Both the sales engineer and the customer may believe that a properly bandwidth-limited digitally sampled signal can be reproduced almost perfectly, right up to the Nyquist limit.

Experienced users generally expect to see small amounts of waveform distortion on a digitally driven display device. The commonly believed causes are:

(a) The hardware/software performs some linear or spline-fit interpolation between actual sample points and forms a quasi-continuous waveform representing the action of the System Under Test. Neither of these interpolation methods make use of the Sampling Integral. So therefore some slope errors will occur.

(b) The display device is often a raster-based system, either because the print-head mechanism is printing dots horizontally across the paper, or because a cathode-ray tube is displaying the waveform on the screen by electron beam writing. Therefore, besides the effects of interpolation as in (a), the waveform is distorted slightly by the jumps from raster line to raster line. The edges of these signals may be slightly distorted.

(c) Signals having segments of rapidly rising or falling data, such as square waves in digital circuits, or signals are known to exceed the sampling frequency or the fastest slew capability of the display mechanism. Anytime they are shown they will be distorted.

(d) Natural noise or occasional transients in the signals which are caused by various natural phenomena, such as Johnson noise or flicker noise. Extremely fast transients in the signals are analog in nature and may cause some analog overshoot or ringing in the input device (i.e., oscilloscope probe) used to connect the signal under test to the analog-to-digital converter circuitry. But these effects have an analog origin and little to do with digital sampling.

The effects of the above forms of distortion can be reduced by careful instrument design, using higher speed sampling or finer dot pitch display technology, so that usually the distortions are small enough to be ignored.

Signal reconstruction is usually omitted for one of several reasons. For instrumentation waveform recorders, textbook discussions (ref. 4) show that adequate reconstruction of the original waveform by analog means is theoretically possible and intuitively a requirement. However, the inherently high cost for design and fabrication of a wide-band adaptive analog reconstruction filter built into the system output would add a bottom-line price increase and discourage customers. And, there is sometimes something counterintuitive about having an analog output device in a digital waveform display system.

On the other hand, waveform reconstruction by a digital processor using the discrete sampling integral also requires too much cost for the computation rate required, and often (as in real-time TV images), there is no reasonably priced processor capable of performing the myriad of calculations in microseconds. More will be said about the processing problem later in this document.

3. UNCOVERING A NEW DISTORTION

A marketing comparison study was performed with the goal of documenting the sampling and display accuracies of some competitive commercial digital strip chart recorders and oscilloscopes. Built into one of the subject strip chart recorders, above the paper output, was a linear bar of light-emitting diodes which assisted in location and alignment of the trace being plotted on the paper. The recorder input was driven by a high-quality analog signal generator. During a wide frequency sweep being used to generate a nearly solid waveform stripe which would document the "flat response of the system," a "beat" in the waveform amplitude shown in the light bar occurred near a certain single frequency on the signal generator. In order to further investigate this, the signal generator sine-wave frequency was "tuned" to reduce the frequency of the beat, and the chart printing drive was enabled. Next, the paper print speed was set low enough that the sine-wave waveform produced a nearly solid swath of printing. Ideally, if the strip chart recorder had shown the "flat" response up to near the Nyquist limit, as claimed in the specifications for the instrument, the envelope of the printed waveform would have been a nearly solid stripe of printing with a straight top and a straight bottom. But with the test waveform, there was an obvious envelope modulation, reproduced here in figure 2. We noted the obvious depth of the notch in the modulation, which was an immediate subject of concern since the signal generator and strip-chart recorder seemed both to be in otherwise excellent operating condition.

Further experimentation showed that the effect was not a fluke. There was a whole family of frequencies exhibiting these deviations from the expected flat trace (seen in fig. 1). Each test required finding a new "beat" and

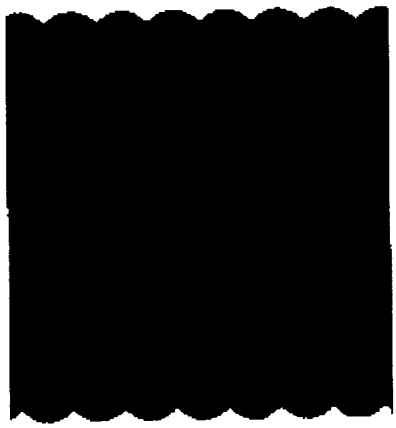

Figure 2.—An actual waveform captured with a commercial digital thermal strip chart recorder. The very black solid thermal printing on the paper was scanned into a digital photograph form for this reproduction. The sample rate in the strip chart recorder was measured as 12170 Hz. The signal was from a sinewave signal generator set to 1367 Hz. Thus the ratio of signal frequency to sample rate was 0.1130649, or almost exactly 6/53. The problem in this example is that the envelope of the waveform (top and bottom edges) should be almost exactly flat. Even allowing for some amount of digital sampling "noise" the envelope was not expected to have the well-defined cusping. The cause of this effect became a study topic.

carefully adjusting the sine-wave generator frequency control to center on each one. Each frequency where a "resonance" occurred was recorded, and the strip chart record was removed and saved, although at the time the cause of the distortions was not obvious. Centering on the "beat" frequency required careful tuning. As the signal generator frequency was very slowly swept past the "center frequency," on each side of the center frequency, the envelope modulation would increase in beats per second and the depth of the notch in the dark stripe would decrease until it faded away, or was replaced by a whole new set of beat events at another critical frequency.

Obviously a simple sine wave of constant frequency has one frequency, and is therefore inherently band-limited to that frequency. Fourier theory shows that single sine-wave has one signal frequency and requires no harmonics to exist. The literature says that, providing that the Nyquist limit is not exceeded, there is no way for the sine-wave signal to produce aliased waveforms. So this resonance effect and the characteristic notching of the waveform modulation envelope were disconcerting.

Ultimately, this test was repeated with a set of different devices, using a digital storage oscilloscope made by an unrelated manufacturer, with input from a different sine-wave generator also made by a different unrelated manufacturer. The results were fundamentally and exactly repeatable. This was not something happening only to strip-chart recorders. A sample of the actual oscilloscope output is displayed in figure 3. These modulations seem to point to a common cause independent of any particular hardware design or manufacturer.

Over time, some study of the technique and frequencies involved resulted in the development of a mathematical model for what had happened. In this model, an arch-typical amount of envelope modulation occurs at a signal frequency which is 6/53 (0.113207...) of the sample rate, or roughly 1/9 of the sample rate. This result is always reproducible with recording equipment which is in good operating condition. In fact, as will be shown, this result is reproducible on a computer spreadsheet, as in figure 4. The reader is invited to perform an independent trial of this experiment. The results will be the same.

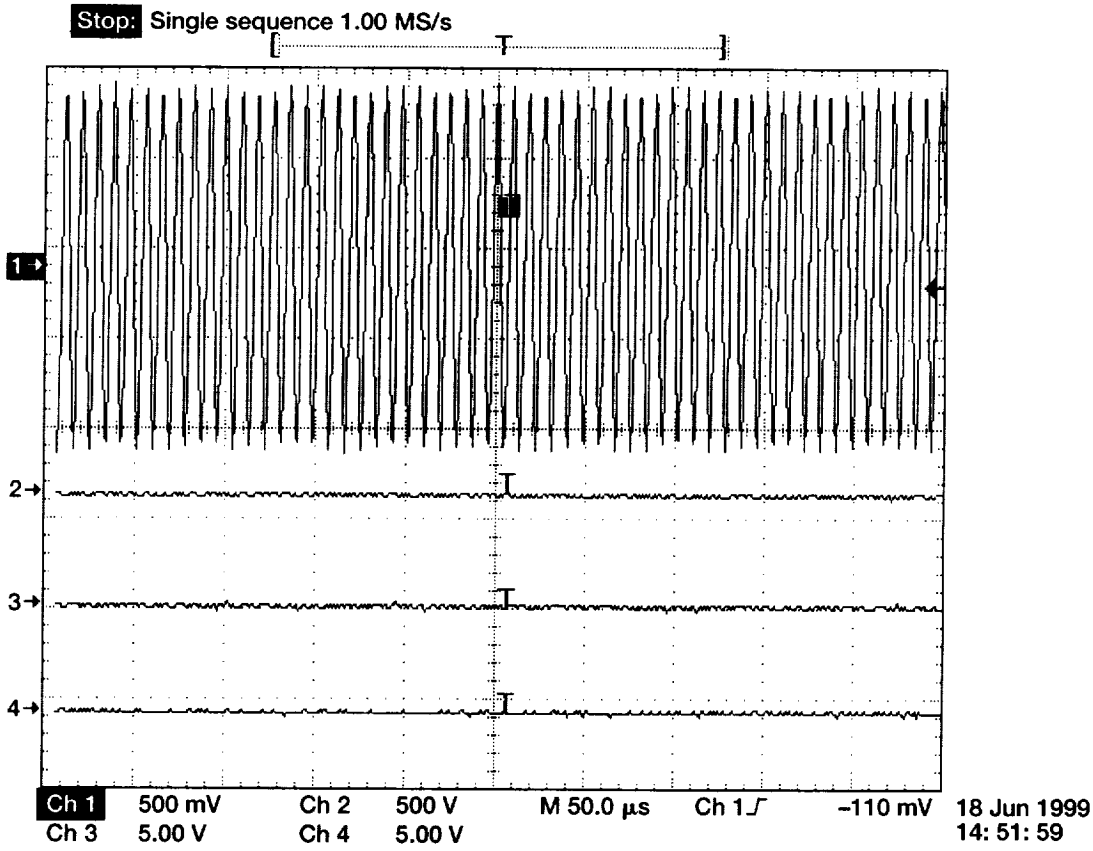

Figure 3.—An actual waveform captured with a commercial digital storage oscilloscope and saved as a graphics file. The sample rate is shown. The sine wave input frequency on Channel 1 was 113.20754 kHz from a commercial 1 Volt peak-to-peak synthesized waveform generator.

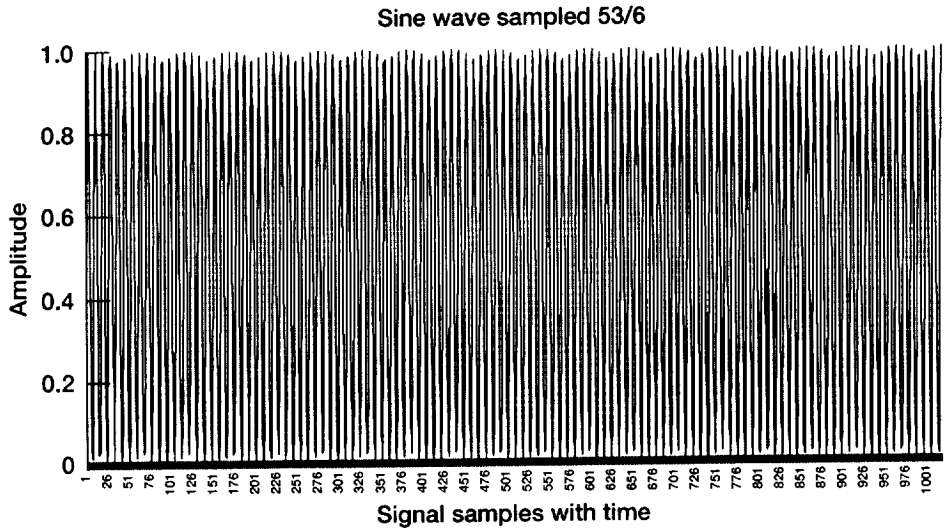

Figure 4.—The condition of sampling a sine wave at any given ratio between signal frequency and sample rate can be reproduced in Microsoft Excel™ without having to run a test with a strip chart recorder and actual electronic instrumentation or Analog-to-Digital Converter. Here the special case of $\frac{m}{n} = \frac{6}{53}$ is synthesized from calculated data and plotted in a chart form.

Since the time of the above experiments, various texts and papers on sampling theory, Moiré patterns, and the like have been researched, without ever finding a similar underlying model. Because of concern that distortion like this can creep into important scientific waveform recording, and even into video image reproduction, the model below is proposed for others to understand and apply in their work.

By now the reader should understand that because signal reconstruction is omitted from the digital display device (or software equivalent), distortion results. This distortion appears to be a modulation of the envelope of the signal, quite often containing sufficient "percentage of modulation" to have a serious affect on the displayed results. Now let us learn why it occurs.

3.1 Development of the Model

A constant frequency and constant amplitude sine-wave waveform is most certainly a bandwidth-limited waveform, expressed as

$$f(t) = \sin(\omega t)$$

where

$$\omega = 2\pi f.$$

Assume that this waveform is sampled such that

$$f \le f_c = \frac{1}{2} f_s \tag{2}$$

where f_c is the Nyquist frequency and f_s is the sample rate. Then we know that the Nyquist limit is not being exceeded. Now we simplify equation (2) as

$$f \le \frac{1}{2} f_s. \tag{3}$$

We then generalize by defining arbitrary positive real integers m and n such that

$$f \leq \frac{m}{n} f_s. \tag{4}$$

with the constraint that

$$m \leq \frac{1}{2} n \tag{5}$$

or, as a redefinition of the Nyquist limit,

$$2m \leq n. \tag{6}$$

Now we apply this to the above special case where $m = 6$ and $n = 53$. The number 53 is a rather uninteresting and seldom mentioned prime number, while 6 is the product of two primes, 2 and 3. We note that 6 and 53 are mutually prime (they have no common divisor other than 1). Whenever these two numbers are involved in a repeating system, there is a cycle which is 6 times 53, or 318, long.

3.2 The Wheel Analogy of Sampling System

Assume the example of two wheels with tires rolling in parallel at a constant translational velocity down a straight track. The outside diameters of the tires are in the ratio of 6 to 53, as in figure 5.

A mark on each tire cycles up and down in the y-axis so as to form a sinusoid[2]

$$y = \sin(\omega t)$$

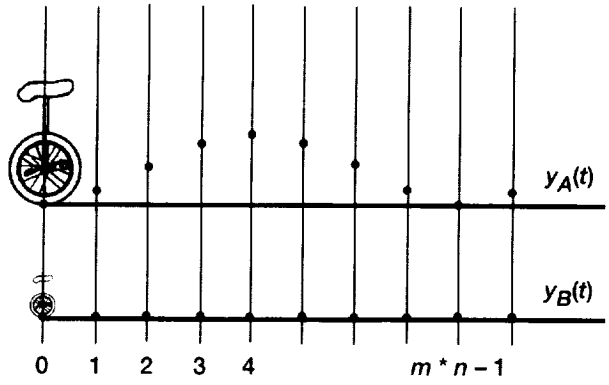

Figure 5.—Position of marks on tires A and B as a function of time, with the elevation y of the marks shown every time the elevation of the mark on Tire B is equal to zero. Effectively, those points occur at the sample times as enumerated. The elevation of the marks on both tires will be zero again at time $t = m * n - 1$ when exactly m rotations of Tire A and n rotations of Tire B have occurred. For most products $m * n$, the number of points to show is very large and therefore this drawing is NOT TO SCALE.

[2]This equation can be derived by applying the parametric equation of motion for a point on a circle of radius r as the circle rolls to the right at unit velocity, with the locus of points forming a cycloid. If the equation of the circle is expressed as $x^2+y^2=r^2$, then the y-axis position of the point is $y=r(1-\cos(t))$ for all t. For a unit circle moving at unit velocity, $dy/dt = \sin(t)$.

Due to the ratio of diameters, and thus circumferences, the larger tire will only have rolled 48/53 of one turn after the smaller tire has made 8 whole turns. After one more turn of the smaller tire, the larger tire will have rolled 54/53 of one turn. It will be impossible after only one turn of the larger tire to have its mark strike the track at same time as the mark on the smaller tire. After the second 8 cycles of the smaller tire, the scenario will be approximately repeated, with an additional fraction of a turn remaining (or exceeded) on the larger tire.

But (since $6 \times 53 = 318$) after a total elapsed time 318, both tire marks will touch the track again at exactly the same moment. Thereafter, if the track is long enough, the cycle is repeated every 318 time units, since 6 and 53 are mutually prime numbers. The distance cannot be smaller because 53 is a prime number, and 6 is a product of the two primes 2 and 3 and cannot evenly divide into 53.

Consider now a system of sampling, wherein the height of the mark on the large tire is logged every time the mark on the smaller tire contacts the track. If charted for thousands of unit intervals, the resulting sine-wave waveform will have impressed on it a modulation or envelope ripple caused by the repeated misses on the peak amplitude of the mark on the larger tire. (It is left as a challenge to the reader to use a charting program or spreadsheet on a modern computer and reproduce the essentials of fig. 4.)

Once the existence of the modulation is recognized, the implications for the world of one-dimensional signal sampling, control system design, telephony, and so forth, are immediately recognized. These implications become now the subject of the remainder of this paper.

3.3 Examples

Many authors, in discussing sampled data systems, fall back on the Nyquist criterion for determining when a signal can be usefully sampled and when it cannot. For the above wheel scenario, the Nyquist criterion implies that the circumference of the larger tire should be at least twice the circumference of the smaller tire.

The initial choices of 6 and 53 for m and n in equation (4) happen to yield a large, obvious modulation. The reasoning applies to the general case of any two mutually prime integer numerators and denominators constrained as in equations (4) or (5). For various m and n, the modulation will still be present in larger or smaller amounts. (A method to calculate the percentage of modulation is shown below.)

Of course, if m is exactly one-half of n, the Nyquist frequency is being sampled and the output is useless. For m/n = 1/3, 1/4, etc. similar ugly results are obtained. But there are dozens of other highly visual possibilities in the range where

$$\frac{1}{20} < \frac{m}{n} < \frac{1}{2}.$$

The reader has already been encouraged to try the above model with a spreadsheet program. It is even more dramatic to set up a live experiment with an oscilloscope or digital strip chart recorder and a sine wave signal generator. The experimental apparatus will need a recorder/oscilloscope having a thermally stable constant sample rate clock. The signal generator should be thermally stable as well and preferably have a digital frequency readout and adjustment means, because the zones of visible modulation are quite narrowly located around the m/n nodes. Of course, for this paper the experiment has already been completed and actual sampled waveform plots have been supplied.

Some readers may not have access to spreadsheet software. For those cases, Listing 1 supplies a dynamic computer program in BASIC, so that a reader having only BASIC can still study the problem. The BASIC program creates a real-time display on a PC having GWBASIC™ or QBASIC™ or similar BASIC, with the ability to set the screen to CGA mode (for clarity) for plotting the waveform on a dark background using the LINE() function. The operator can input the frequencies of interest, in order to experiment in real time with the effect.

Finally, the reader should be aware that this problem occurs in all sampling systems where there is a constant frequency sample rate or a constant spatial distance between samples. In statistics, the problem could occur any time data is collected or analyzed on a cyclical basis. For instance, assume a key economic indicator is analyzed monthly by an economist, and the results are published in the economics literature. The economist maintains that the indicator has a hitherto unrecognized cycle, a cycle which is almost 106 months long. But 106 months of a monthly published statistic could be related by 106/12 or 53/6. An obvious distortion in the data could occur having nothing at all to do with a real cycle or a real fact.

LISTING 1.

```
10 REM PRogram to print samples along sine wave
20 CLEAR : CLOSE : CLS : KEY OFF
40 PI = 3.1415926536#
50 FS = 1000!
60 INPUT "What is the sample rate ratio"; RAT
70 FP = 1000!
80 TN = 2 * PI / RAT
90 Y = 0!
100 Z = 0!
110 TPI = 1! / FP
120 PT = TPI
130 CLS : SCREEN 1
140 COLOR 0, 1
150 NN% = 0
160 FOR N = 1 TO 10000 STEP 1
170 T = N / FS
180 X = SIN(TN * N)
190 IF Y < X THEN Y = X
200 IF Z > X THEN Z = X
210 IF (T < PT) GOTO 320
220 YP% = FIX(79! * Y) + 100
230 ZP% = FIX(79! * Z) + 100
240 NN% = NN% + 1
250 IF NN% > 319 THEN NN% = 0
260 LINE (NN%, 0)-(NN%, 199), 0
270 LINE (NN%, ZP%)-(NN%, YP%), 2
280 REM PRINT NN%,ZP%,YP%
290 PT = PT + TPI
300 Y = X
310 Z = X
320 IF (INKEY$ <> "") GOTO 340
330 NEXT N
340 CLS : SCREEN 2: SCREEN 0, 0, 0
350 END
```

3.4 A Video (Television) Example

A very common example of the problem occurs in television broadcasting, a medium which most people view more often than they view one-dimensional waveform plots. A television receiver is a system displaying dynamically changing two-dimensional waveforms. To the eye, the two-dimensional waveform comes in sets of interleaved raster lines forming static images which are displayed for approximately one-thirtieth of a second each. As we know, the stream of images occurring at nominally 30 frames/sec (in the United States) is rapid enough to fool the human eye/brain system into seeing a continuous moving picture.

Each video frame is a static image composed of a set quantity of raster lines "painted" across the screen by the electron guns inside the cathode ray tube. As the electron guns and yoke magnets steer the painting of a raster line, the shadow mask interrupts the beams in order to break the display up into areas of red, blue, and green "dots." The shadow mask is therefore a spatial sampling system having a spatial frequency. The electron guns and yoke magnets inside the television's cathode-ray tube compartment are driven by smooth analog signals derived from the demodulation of the radio-frequency television carrier signal. Therefore there is a time element to the display.

Another type of sampling of the video image occurs because of the vertical arrangement of the horizontal raster lines which form the images. Each image is effectively sliced into hundreds of horizontal strips by the order in which the electron guns paint the raster lines. The separation of these horizontal strips at a constant pitch forms another spatial frequency for sampling.

The sampling problem begins when the video image happens to contain a fine-grained repetitive pattern, so that the spatial pitch of the image pattern, another spatial frequency, will be exactly a fractional portion of the shadow mask sampling frequency, or the vertical raster pitch

Repetitive patterns occur in all sorts of ways, such as views of venetian blinds, clothing with patterns, the flag of the United States, etc. A good example of such an image having a large area with a rapid spatial frequency is that of a person wearing a patterned or finely striped tie or jacket. In figure 6, a person was wearing a jacket with a tight hound's tooth or herringbone twill weave. The weave was very visible and very obvious on the screen, so it did not fit the classic description of an alias caused by having a spatial frequency higher than the sampling frequency (dot pitch) of the cathode ray tube screen. Most people have seen these visual effects and at times the effects can be quite annoying. Most people can find an example of them at least once per hour of television viewing. The manufacturers of VCRs and TVs call these effects Moiré patterns and claim their hardware minimizes the effects of these patterns.

In order to further show that distortions in video images can occur if an object in the image has a pattern which is harmonically related to the spatial frequency of the display device, a 512×484 image was synthetically created with a C program (Listing 2). Each of the 484 raster lines of the image are identical, and result from calculating and repeating a raster line containing 512 samples of a sine-wave waveform having a frequency which is (you

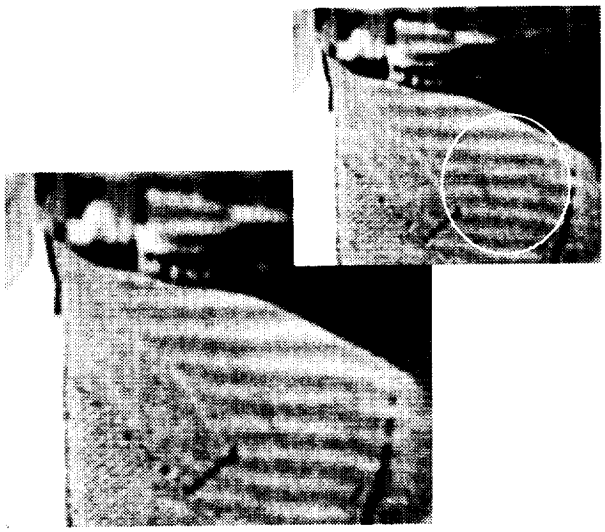

Figure 6.—This image was photographed off the screen of a color television with a 35 mm film camera. The announcer's jacket consisted of a twill tweed pattern that was obviously visible on the television screen. Therefore, the spatial frequency of the cloth weave pattern was lower than the TV screen dot pitch Nyquist frequency. Then, as the announcer moved, the aliasing effect (inset circle) became obvious. To make this image of use in this document, because scanners and digital laser printers can create secondary aliases in this situation, an 8x10 print of the negative was ordered. The print was then scanned on a high resolution scanner. The resulting image was computer cropped to show the portion of the image containing a series of very obvious aliasing patterns in the announcer's left shoulder area. It should be noted that this figure is neither a proof that the patterns are caused only by the Moire' effect, nor that they are specifically caused by sub-Nyquist distortions.

Listing 2. The C source code for making a raw image file which exhibits alias bars (Figure 7).

```
/* RASTER.C - Create basic raw image raster 512x484 monochrome

Usage:    RASTER  outfile   divisor  [> textdatafile]

*/

#include <fcntl.h>
#include <sys\types.h>
#include <sys\stat.h>
#include <io.h>
#include <stdlib.h>
#include <stdio.h>
#include <signal.h>
#include <math.h>

#define PI 3.14159265359

unsigned char buff[512] ;
int fh ;

/*  PROTOTYPES */

int makefile( char * );
int makeline( double );
int copyline( void );
static int ctrlc_handler(int);
void cleanup ( void );

/*--------------------------*/

main(int argc, char *argv[] )
{
double incrr ;

    atexit(cleanup);

    if ( signal(SIGINT, ctrlc_handler) == SIG_ERR )
      exit(2);

    if (argc <= 2 )
    {
      exit(2) ;
    }

    incrr = atof( argv[2] ) ;
    makefile(argv[1]);
    makeline( incrr );
    copyline();
    exit(0);
}

int makefile(char *filename)
{
   if ((fh = open( filename, O_RDWR | O_CREAT | O_EXCL | O_BINARY, \
                              S_IREAD | S_IWRITE ) ) == -1 )
       {
          printf("Error opening file %s\n ", filename);
          exit(1);
       }
       return(0) ;
}

int makeline(double incr)
{
```

```
int i ;

    for(i = 0; i < 512 ; i++ )
    {
        buff[i] = (unsigned char)(127 * (1+sin((2*PI*i)/(incr) ) ) ) ;
    }
}

int copyline()
{
int j ;

   for(j = 0 ;  j < 484 ; j++ )
   {
     write(fh,buff,512) ;
   }
   for(j = 0; j < 512 ; j++)
   {
      printf("%d\n", buff[j] ) ;
   }
}

void cleanup (void)
{
  if ( fh != -1 )
   close(fh);
}

int ctrlc_handler(int sig)
{
   exit(1);
}
```

guessed it) 6/53 of the sample rate. This image (fig. 7) shows the symptoms. (Look for a few extra dark vertical bands.) Since an image is "virtual," one could assume that this image is actually a monochrome frame-grabbed video image of a rectangular panel on a wall in a room, and the panel happens to have a sine-wave fringed pattern like the image, but without the distortion. Once the camera taking the image is positioned correctly, the lens and charge-coupled device (CCD) sensor in the camera can create this effect if the fringes "beat" with the 512 dot spatial frequency of either the camera or the digital frame grabber.

4. GRAPHING THE FIELD OF POSSIBILITIES

The interesting ratios m/n form a mathematical field of fractions where the numerators and denominators are mutually prime integers or products of mutually prime integers. When graphed in Cartesian coordinates such that n is the ordinate and m is the abscissa, and little crosses (×) or nodes mark each $\{m,n\}$ coordinate, the resulting graph resembles a minefield of nodes (fig. 8). The nodes representing the Nyquist frequency can be connected as a line of slope equal to 2 crossing the origin at {0,0}, hereby named the Nyquist line. For clarity, figure 8 has been drawn so that nodes on the far side of the Nyquist criterion ($m > n/2$) have been excluded.

In figure 8, the two regions, one on each side of the Nyquist line, are labeled as "I" and "II" for a particular reason. Region I is the region where classic signal alias is defined. Region II defines the domain where sub-Nyquist distortions, the subject of this paper, are defined. Most digital sampling applications occur in Region II. It is where billions of dollars are invested in technology depending on accurate sampling. Remember that in most of these applications, the Sampling Integral is never included as part of the waveform reproduction or display means.

The reader by now should realize that all of the above are special cases. Normally the sampled signal is not sinusoidal, but of course the signal is assumed to be band-limited. Also, the ratio between a signal frequency and the sampling frequency is for all practical purposes a "random" ratio somewhere in Region II, hopefully and we assume

The calculation proceeds as follows. From the graphed figures in this paper (or the wheel analogy), it is clear that the overall period of the modulation error is $m*n$. It is also clear that the time $m*n$ represents the entire cycle of the modulation error, and therefore, is the time or horizontal distance from one peak to the next peak of the modulation ripple. The number of samples from peak to valley of the modulation is then equal to

$$\frac{m*n}{2}. \tag{8}$$

Starting from t_0, after approximately one full cycle of the signal at rate m, exactly n samples will have been acquired.

It is useful now to convert from time units to angular vector relationships, to radians of revolution, and treat the next few equations in radians rather than time. A full cycle of the signal is 2π radians, and each sample will be spaced approximately $2\pi/n$ radians apart. At the end of n samples, there will be a residual angle left over, plus or minus, which must be applied as a debt or credit to the following signal cycle of 2π radians. This error slowly accumulates over many periods to reach a maximum where the valley of the ripple has the greatest depth. Therefore, the total length from peak to valley is, converted from above,

$$(\text{one} - \text{half \# of samples}) * (\text{distance between samples}) = \left(\frac{m*n}{2}\right) * \left(\frac{2\pi}{n}\right) \text{radians}. \tag{9}$$

There is a temptation to simplify equation (9), but let's not do it yet. We need to find the left over phase error at the valley, which is calculated by calculating the residual phase error modulo m:

$$\theta = \left[\left(\frac{m*n}{2}\right)\left(\frac{2\pi}{n}\right)\right] \bmod(m) = \left(\frac{2\pi}{n}\right)\left[\left(\frac{m*n}{2}\right) \bmod(m)\right] \tag{10}$$

Now, we need to make use of the fact that m and n are mutually prime integers. Since, per equation (6), m is n less than m, then m is the only one of the two integers which can take on the value 2. Therefore, n, being greater than two and prime, will always be odd, and $n/2$ is also not a whole number. But, if n is odd, then $(n-1)$ is even, and $(n-1)/2$ is a whole number. Then,

$$\begin{aligned}\left(\frac{m*n}{2}\right) \bmod(m) &= \left[(n-1+1)*\frac{m}{2}\right] \bmod(m) = \left[\left((n-1)*\frac{m}{2}\right)+\left(\frac{m}{2}\right)\right] \bmod(m) \\ &= \left[m\left(\frac{n-1}{2}\right)\right] \bmod(m) + \left(\frac{m}{2}\right) \bmod(m) \\ &= 0 + \frac{m}{2}\end{aligned} \tag{12}$$

and therefore, the total phase error is

$$\theta = \left(\frac{m}{2}\right)\left(\frac{2\pi}{n}\right) = \pi\left(\frac{m}{n}\right). \tag{13}$$

The normalized full-scale amplitude of the signal in the valley of the notch, a_v, is then:

$$a_v = \cos(\theta) = \cos\left(\frac{\pi m}{n}\right). \tag{14}$$

$$a_v = 0.937\ldots$$

which amounts to a 6 percent peak error!

Given that amount of error amplitude, we are curious to evaluate when the distortion error from a Region II node is smaller than a certain percentage of full scale. More important is to determine when the error is small enough to be inconsequential. In an n-bit sampling system, any error smaller than the smallest step sample size will not be resolvable. Referring to figure 9, for the normalized full-scale error value ε such that:

$$\varepsilon = 1 - a_v = 1 - \cos\left(\pi * \frac{m}{n}\right) \tag{15}$$

when applied to an 8 bit sampling system, gives an error ε smaller than one bit which is

$$1 / 256 = 0.4 \text{ percent.}$$

Solving for the ratio *m/n* in equation (15)

$$\frac{m}{n} < \frac{1}{\pi} * \arccos(1 - \varepsilon)$$

and applying equation (16) to the above error of 0.4 percent,

$$\frac{m}{n} < 0.028.$$

Therefore, for a telephone line channel sampled at 8000 times/sec (a common telephone industry sample rate for a private line) sampled with an 8-bit analog-to-digital converter, distortion is negligible below 224 Hz, a frequency lower than the musical note middle C.

Similarly, for a 10 bit waveform recorder (1000 points across the print head) sampling at 10,000 Hz, a manufacturer might claim that the instrument is "flat to 5000 Hz," which is the Nyquist frequency. So to display no Region II modulation above one dot peak error (0.1 percent) at full scale requires having no signals exceeding 142 Hz! Typically, instrument manufacturers will be reluctant to admit this constraint exists because, through no

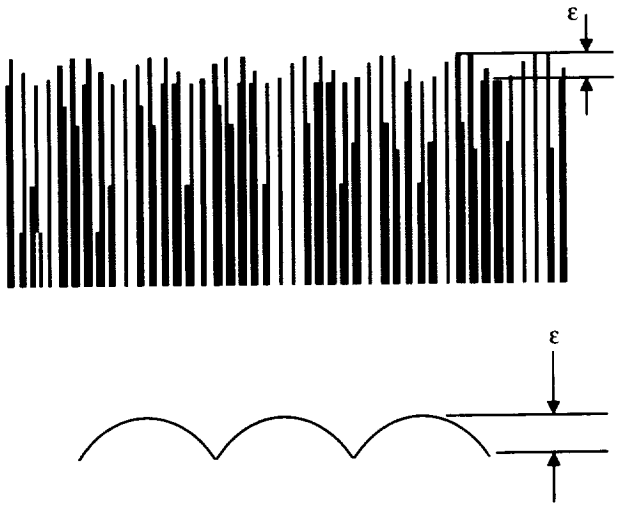

Figure 9.—The maximum percentage modulation error ε depends on determining the distribution of samples near the envelope peak, as in equation 15.

fault of their own, this constraint would make the specifications for their instrumentation equipment appear to be rather poor.

4.2 More About the Nyquist Theorem

Industry literature is rife with descriptions of how aliasing occurs whenever a sampled signal exceeds the Nyquist limit. Nothing has been found in the literature which describes significant signal peak modulation distortions caused by sampling errors for signal frequencies far smaller than the Nyquist limit. This study has shown that, in spite of assumptions based on recurrent descriptions in the literature of aliasing errors caused only by violations of the Nyquist limit, a distortion error can occur at frequencies far below the Nyquist limit.

One might ask an almost obvious question at this point. Why not just use the formulas above and "correct" the distortion effects by putting in a variable amplification factor to remove the dip in the waveform? The answer to that is two-fold. First, the locations of the distortions are very much dependent upon the phase between the signal and the sampling frequency. There is no way a priori to determine the relative phase shift. That is, on which sample will the waveform start to dip, given that in the application we do not even know the waveform's frequency or phase or amplitude until after the waveform distortions occur? Second, we know that a typical waveform is not a perfect sine-wave, but is usually a composite of many waveforms, per Fourier theory, with some added system noise from the "real world." Therefore the offending distortion, and only the offending distortion, would have to be isolated first in order to perform the correction. It would be far easier to implement the reconstruction via the Sampling Integral and let all the corrections be done with a consistent provable algorithm.

Unfortunately, past attempts to perform waveform reconstruction on a computer, given unlimited computation time, using both FORTRAN and MATLAB™ has shown no feasible solution to the reconstruction problem. Given a finite sequence of samples, we have been able to apparently reconstruct a sine-wave without modulation distortion. But the very same algorithms fail miserably to reconstruct square-waves, and we can infer that other waveforms such as sawtooth waveforms would also show severe reconstruction errors. These errors have been known for years as resulting from the Gibbs phenomenon and there have been numerous studies in that area (ref. 7).

A conjecture will now be made. Aliasing errors are not just confined to Region I on figure 8. It is also possible to consider that the envelope distortions which occur in Region II are really a form of "sub-Nyquist" alias. Therefore, in this report, Region I aliases could be named "Type I aliases." Similarly, Region II modulations could be named "Type II aliases."

5. CONCLUSIONS

In waveforms, statistical data, and images, sample aliasing due to system coincidences can occur in one of two domains as defined by the Nyquist frequency line in figure 8. The two domains have the characteristic that:

(a) Classical aliasing in Region I occurs anytime where and are defined as in equation (4), i.e., and are real numbers such that $n \geq 2m$.

(b) Region II distortion occurs for any positive real integers m and n where

$$1 \leq m \leq \frac{n}{2}$$

and

$$n \geq 2$$

and m and n are mutually prime. The integers m and n separately can be products of prime numbers without affecting these conclusions. These mutually prime integers and products m and n form a mathematical field of possible trouble points, or nodes. Systems relying on sampling for gathering data obviously should avoid taking data in the condition of Region I, i.e., when the Nyquist limit is exceeded. But system designers should also avoid Region II nodes whenever possible if they are implementing systems which do not perform adequate waveform reconstruction before presenting plots which display critical information.

This report has shown that a new domain of distortion, in Region II, has subtle implications for the fabrication of systems using digital waveform sampling. Except for television, where the effects of swimming color bands are obvious and even obnoxious, there has not been a great deal of attention focused on this type of alias. However, in the future, engineers and statisticians should determine what impact the Region II distortion may have their data before drawing conclusions.

Finally, in this report, no detailed analysis has been done to see if the modulation effects around a Region II node result in extra peaks in the power spectrum indicating signal power is aliased into undesirable frequencies. No claim is made that the Region II distortions result in real signal power being lost from the sampled signal. However, we are concerned that in rare cases the envelope distortions could be interpreted as modulation and cause serious consequences in error detection systems and feedback control systems.

6. REFERENCES

1. C.E. Shannon: "Communication in the Presence of Noise," *Proc. IRE*, Vol. 37, pp. 10–21, Jan. 1949.
2. Ahmed I. Zayed: "Advances in Shannon's Sampling Theory," CRC Press, New York, 1993.
3. Chi-Tsong Chen: "One-Dimensional Digital Signal Processing," Marcel Dekker, Inc., 1979, pp. 69–83.
4. Mischa Schwartz: "Information, Transmission, Modulation, and Noise," McGraw-Hill, 1959, Chapter 4.
5. "Data Acquisition and Conversion Handbook," edited by Eugene L. Zuch, ca. 1978, published by Datel-Intersil Corp., p. 236.
6. Abdul J. Jerri: "The Shannon Sampling Theorem—Its Various Extensions and Applications: A Tutorial Review," *Proc. IEEE*, Vol. 65, No. 11, Nov. 1977, pp. 1565–1596.
7. D. Gottlieb, C.-W. Shu, A. Solomonoff, and H. Vandeven, "On the Gibbs Phenomenon I: recovering exponential accuracy from the Fourier partial sum of a nonperiodic analytic function," J. Comput. Appl. Math., v43, 1992, pp. 81–92. (See also subsequent papers II, III, and IV.)
8. P. Mertz and F. Gray: "A Theory of Scanning and Its Relation to the Characteristics of the Transmitted Signal in Telephotography and Television," *The Bell System Technical Journal*, Vol. 13, July 1934, pp. 464–515 (in "Graphical and Binary Image Processing and Applications," edited by J.C. Stoffel, Artech House, 1982, pp. 5–56.)

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	October 2000	Technical Memorandum

4. TITLE AND SUBTITLE

Sub-Nyquist Distortions in Sampled Data, Waveform Recording, and Video Imaging

5. FUNDING NUMBERS

WU–940–30–09–21

6. AUTHOR(S)

Glenn L. Williams

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

National Aeronautics and Space Administration
John H. Glenn Research Center at Lewis Field
Cleveland, Ohio 44135–3191

8. PERFORMING ORGANIZATION REPORT NUMBER

E–12437

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

National Aeronautics and Space Administration
Washington, DC 20546–0001

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

NASA TM—2000-210381

11. SUPPLEMENTARY NOTES

Responsible person, Glenn L. Williams, organization code 7715, 216–433–2389.

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Unclassified - Unlimited
Subject Categories: 32, 35, and 65 Distribution: Nonstandard

This publication is available from the NASA Center for AeroSpace Information, 301–621–0390.

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 words)***

Investigations of aliasing effects in digital waveform sampling have revealed the existence of a mathematical field and a pseudo-alias domain lying to the left of a "Nyquist line" in a plane defining the boundary between two domains of sampling. To the right of the line lies the classic alias domain. For signals band-limited below the Nyquist limit, displayed output may show a false modulation envelope whenever inadequate signal reconstruction is used before display. The effect occurs whenever the sample rate and the signal frequency are related by ratios of mutually prime integers. For cost and technical reasons, the waveform display devices omit the required reconstruction steps. Belying the principal of a 10:1 sampling ratio being "good enough," this distortion easily occurs in graphed one-dimensional waveforms and two-dimensional images and occurs daily on television.

14. SUBJECT TERMS

Sampling; Nyquist; Alias; Waveform; Analog-to-digital; Digital-to-analog; Shannon; Video; Reconstruction

15. NUMBER OF PAGES

25

16. PRICE CODE

A03

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	